The Effects of Review Bombing & Score Weighting on Movie Recommendations

By: Kai Chang, Oliver Stewart, Sami Jaman, Justin Lederer

INTRODUCTION:

The movie business is an immense industry. According to Box Office Mojo, the US domestic box office gross for 2025 has already exceeded $8.5 billion; each of the two preceding years, the total annual gross fell between $8.5 and $9 billion [Box Office Mojo]. These numbers exclude international box office revenues, money spent to physically or digitally purchase or rent movies, and money spent to watch movies through subscription streaming services. In an environment where various new releases compete for the biggest slice of the box office pie, film studios and researchers are acutely interested in predicting which films will be financially successful and orienting release calendars accordingly.[1]

In the early 2000s, various research explored the effect of critical reviews on film success, and at least one study (Basuroy et al.) found that “each of the first eight weeks, both positive and negative reviews are significantly correlated with box office revenue” [King, 2007; Basuroy et al., 2003]. However, with the rise of the internet, consumers no longer turn to critics like Roger Ebert or Pauline Kael when deciding whether to see a film; instead, word-of-mouth (WOM) information on social media sites and crowd-sourced movie ratings like those from IMDb are among the first information presented to prospective filmgoers. Recent research has examined the effect of online WOM on movies and analyzed online review data, including using sentiment analysis to measure online response to a film [Sharma et al., 2023; Yoon et al., 2017; Oghina et al., 2012].

A fundamental difference between crowdsourced and critical movie evaluation is the inherent manipulability of the former. Online communities, working in concert, can shape online movie scores by leaving masses of (usually negative) reviews to lower a movie’s score, a phenomenon known as “review bombing.” Some recent research has examined this phenomenon, including papers by Schuff et al. (2024) and Tomaselli et al. (2021).

Review bombing is most evident in the scores of films which are perceived as having made “woke” an existing (white and/or male) movie or franchise by rebooting it with women and/or people of color in the main roles (e.g. *Ghostbusters* (2016), *The Little Mermaid* (2023)) and films which address topics like the Israel-Palestine conflict or the Armenian Genocide (e.g. *Israelism* (2023), *The Promise* (2016)). In many cases, although not always, review bombing follows an online backlash, often driven by racism or misogyny. The recent remake of *The Little Mermaid*, for example, drew racist ire for its casting of Black actress Halle Bailey in the lead role, and was consequently review bombed [The Guardian]. When franchises with large, predominantly male online fan bases have produced films with mostly female casts or centering women—for example, the *Ghostbusters* remake, *The Marvels* (2023), and *Star Wars: The Last Jedi* (2017)—those films have likewise been review bombed. In other cases, films have been review bombed simply for portraying events or phenomena that online critics deny or object to seeing

[1] Research by Lash & Zhao (2016) and Lee et al. (2016), among others, has explored the use of machine learning techniques to predict movie success and profitability.

discussed. For example, *The Promise* takes place during and addresses the Armenian Genocide, which is commonly denied by Turkish nationalists, and drew review bombing accordingly, particularly from within Turkey. *Israelism* was review bombed for its portrayal of American Jews learning the facts about Israel's treatment of Palestinians in ways that Jewish institutions had not taught them. Review bombing, then, is frequently driven by bigotry and other violent ideologies, and is specifically designed to punish films and filmmakers who run afoul of those ideologies. Review bombing has become increasingly prevalent in recent years, to the point that it has led sites like IMDb to weight their ratings, leading to sometimes significant gaps between unweighted means and the publicly presented score.[2]

When online reviews can shape film success and by extension future studio decisions, unchecked review bombing has the potential to deter studios and directors from future projects which center or positively portray marginalized groups or address political issues. Ratings from IMDb, which is owned by Amazon, are explicitly displayed on the Amazon Prime streaming service; this represents a direct way in which IMDb ratings contribute to film success, as higher-rated films are more attractive to users and likely to be shown more often. Other movie recommendation algorithms also rely on some form of score to produce recommendations, although these may be more opaque; review bombing, if not addressed, can directly perpetuate representational and political harms by damaging films' performance in recommendation algorithms.

With this in mind, weighting systems like IMDb's represent direct bulwarks against these harms, and how successful it is at identifying and negating review bombing is critical in determining whether those perpetuating the review bombing are successful in damaging the film and its creators. Our project explores review bombing, score weighting, and film recommendations, analyzing which kinds of films are more likely to review bombed, how that relates to their success, how platforms like IMDb use weighting to minimize the effects of review bombing, and more. In addition to our observational analysis, we plan to create our own film recommendation algorithm, and run it using weighted and unweighted ratings to explore how effective score weighting might be in counteracting review bombing, and whether there are any apparent issues with IMDb's weighting methods.

Research Questions:

1. In terms of genre and keywords, what kinds of movies does IMDb's weighting algorithm boost or de-boost?
2. How do different movie rating weightings affect the rate at which movie recommendation algorithms recommend certain movies?

We expect to find that the ratings of movies that have been review bombed contain a bimodal distribution, centering around extremes of 1 and 10 when compared to other movies, suggesting that many users of IMDb rate these movies based on ideology rather than the film itself,

[2] 5.2 vs 7.2, 5.2 vs 6.8, 5.8 vs 7.7, and 6.0 vs 6.1 for each of the previously mentioned movies, respectively. The smaller gap for *The Promise* is due to the occurrence of simultaneous *positive* review bombing, a phenomenon also visible on a smaller scale for some of the other films.

perhaps without even watching it. We also predict that a more diverse cast, especially ones with more underrepresented groups such as an all-women cast, will draw more polarized ratings; which in return will lead to these movies being deboosted in recommendation algorithms.

We will measure the polarization of rating distributions using a version of the polarization score laid out by Esteban and Ray in 1994, which measures the polarization of a distribution by comparing the difference between clusters in the distribution and accounting for the size of those clusters [Esteban and Ray, 1994]. Although we adapted our own version of the score, a version of it was first applied to movie reviews by a FiveThirtyEight analysis by Mehta and Hickey [Mehta and Hickey, 2017].

We expect to find that movies belonging to genres that address real-world issues ("History," "Biography," and "Documentary," for example) will be more likely to have polarized rating distributions and to have average ratings that differ the most across different countries. Examples of individual movies that support this hypothesis mentioned in the introduction are *The Promise* (2016) and *Israelism* (2023) (though *Israelism* will not be included in our initial analysis because it has fewer than 50,000 user votes). We also expect to find that movies with higher proportions of female and non-white stars will be more likely to have polarized rating distributions. Examples of individual movies that support this hypothesis mentioned in the introduction are *Ghostbusters* (2016) and *The Little Mermaid* (2023).

DATA:

We use a dataset scraped from IMDb's website and GraphQL API, cleaned into imdb-cleaned.csv, to examine how user ratings reflect review bombing and polarization. Each row in the dataset represents a single movie and includes variables such as the film's title, genre(s), release year, keywords, rating breakdown (number of votes for scores 1–10), names of the four star actors (as determined by IMDb), and the five countries with the most user votes. Additional variables were created in the data preparation process to better capture polarization and rating behavior, including unweighted_rating, total_votes, rating_diff (the difference between the IMDb weighted rating and the unweighted average), and polarization_score, based on Esteban and Ray's polarization index. The dataset also tracks the difference in average scores between countries (country_ptp) and identifies which countries contributed the highest and lowest ratings (max_country_ptp, min_country_ptp).

For cast demographic information, we match actor names with race/ethnicity and gender datasets from Bamman et al. (2024). Their race/ethnicity dataset is based on crowdworker labels and their gender dataset is from Wikidata.

Some important terms it is necessary to define for our project are as follows: polarization, polarization score, international polarization, weighted rating, unweighted rating. Some of these are fairly simple; the *weighted rating* is a film's basic overall rating provided by IMDb, while the *unweighted rating* is the unweighted mean rating provided on a film's IMDb page. *Polarization* refers to the amount of disagreement in a film's ratings; a film is more polarized if it has a higher *polarization score*. We calculate a film's polarization score using a version of the polarization formula created by Esteban and Ray (1994); the maximum score a film can receive is 1, meaning that exactly half the ratings are 10 and the other half are 1, and the lowest is 0, if all the ratings are identical (Mehta and Hickey (2017) also use a version of this formula, but a slightly different one). *International polarization* of a film between two countries refers to the difference in unweighted mean rating for the two countries; although international polarization did not end up featuring significantly in our work, it could be a fruitful metric for future analysis, so it appears in our dataset regardless. These are some of the basic metrics we will use to identify films that have been review bombed and analyze the extent and nature of the review bombing.

METHODS:

To facilitate our exploration of how review bombing may impact film recommendation systems, we created a basic recommendation algorithm of our own. A common and intuitive use case for recommendation systems is “content-based recommendations,” in which a user viewing one product (or film) is recommended other, similar content [Scientific American].[3] This was a relatively simple use case for us to replicate; unlike recommendations based on general and similar user activity, which require a large body of user data to implement, content-based recommendations require only the kind of film data which we already had as part of our IMDb dataset. They are also ubiquitous, under various names, on streaming services today; Netflix (“More Like This”), Amazon Prime Video (“Related”), Disney+ (“Suggested”), Hulu, and Max (both “You May Also Like”) all show users content-based recommendations whenever they view a film. For these reasons, we pursued a content-based approach in creating our mock recommendation algorithm.

Our recommendation system takes a film and recommends similar movies based on content and rating. For each film in the database, IMDb provides a list of “plot keywords,” words which pertain to or describe the content of the film. To generate recommendations based on a film (the “base film”), our algorithm looks at the top 50 keywords for the base film and identifies potential recommendations by seeing which films have the highest percentage of shared keywords in their own top 50 (or, if a film has less than 50 keywords, in the entire keyword list). We then calculate the percentage of keywords a potential recommendation shares with the base film as a number out of 10, where 0 is no shared keywords and 10 is all keywords in common. For each potential recommendation, that number is then multiplied by that film’s rating to determine a recommendation score designed to capture both film similarity and popularity. (For example, a film with 20/50 shared keywords (4) and a rating of 6.5 would receive a score of 26, while a film with 17/50 shared keywords (3.4) and a rating of 8.1 would receive a score of 27.54). This yields

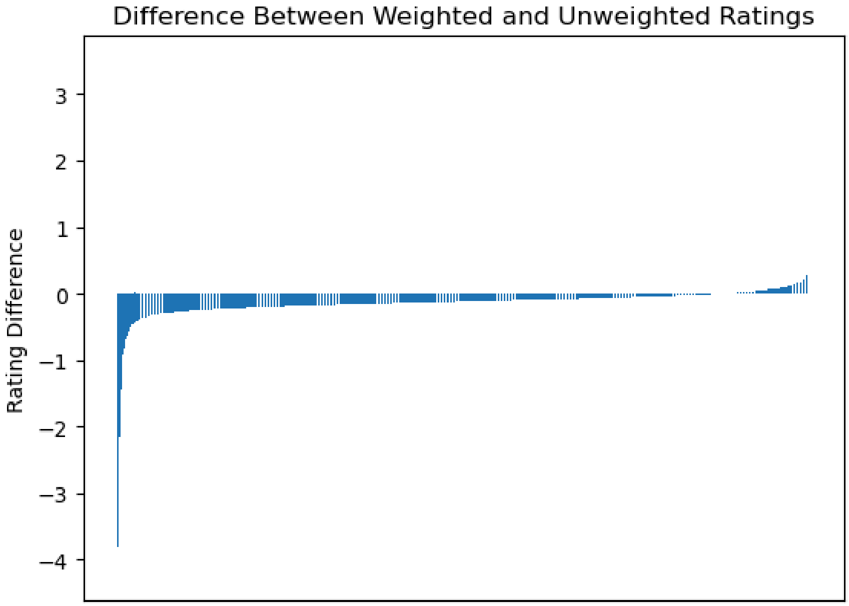
[3] The term content-based filtering can also mean recommendations based on matching content details with general user preferences; our recommendation system covers the more basic, although also real-world relevant, use case of recommendations based on a single film. Recent research by Singla et al., among others, has explored the creation of systems to generate content-based film recommendations; where we use IMDb plot keywords, they assessed film similarity based on factors including genre and doc2vec machine learning analysis of plot summaries [Singla et al., 2020]. Although they created their recommendations based on more complex factors and using more advanced techniques, the aim of capturing film similarity is the same.

recommendation scores between 0 (no shared keywords, any rating) and 100 (identical keywords, 10 rating). Then, the films with the highest recommendation scores with regard to the base film are outputted as recommendations.

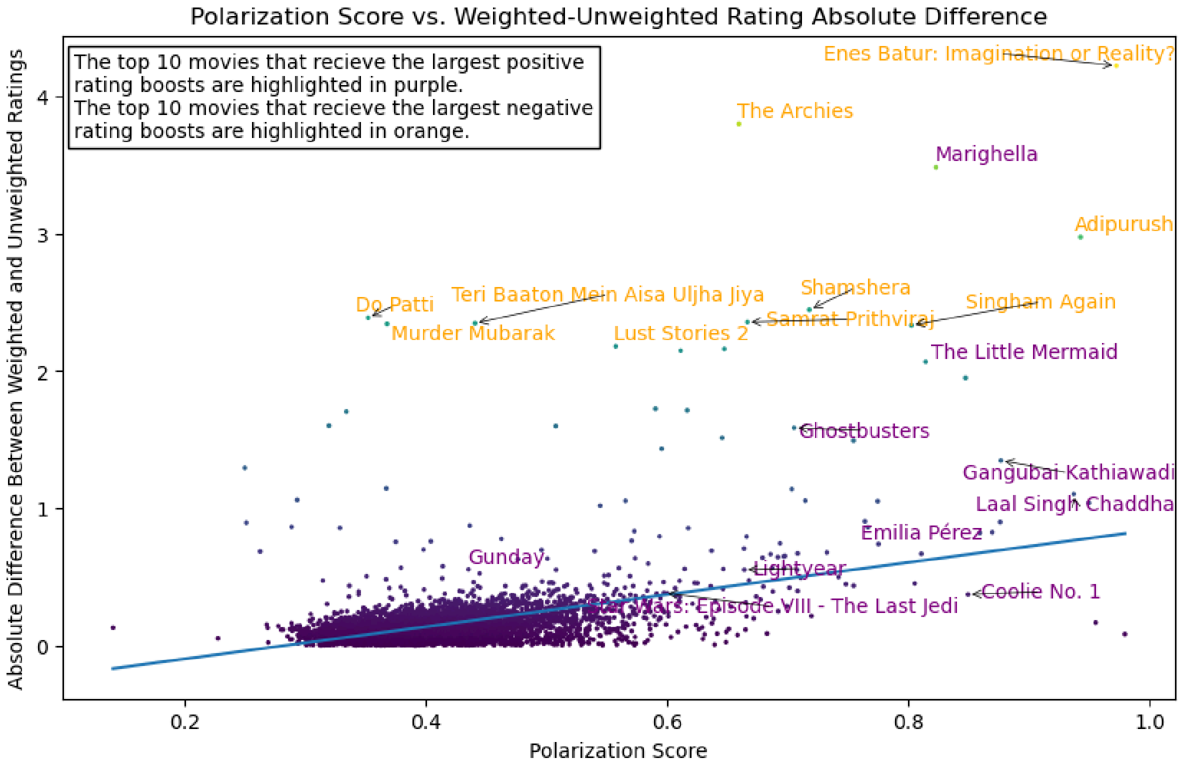
By running the recommendation system with both weighted and unweighted ratings and comparing the difference in results, we aim to illuminate the potential impact of score weighting on recommendation systems. Although our system represents a simplified version of what an actual streaming service's recommendation system might look like, we believe that it should allow for worthwhile analysis of review bombing and score weighting's respective potential effects on recommendation.

RESULTS:

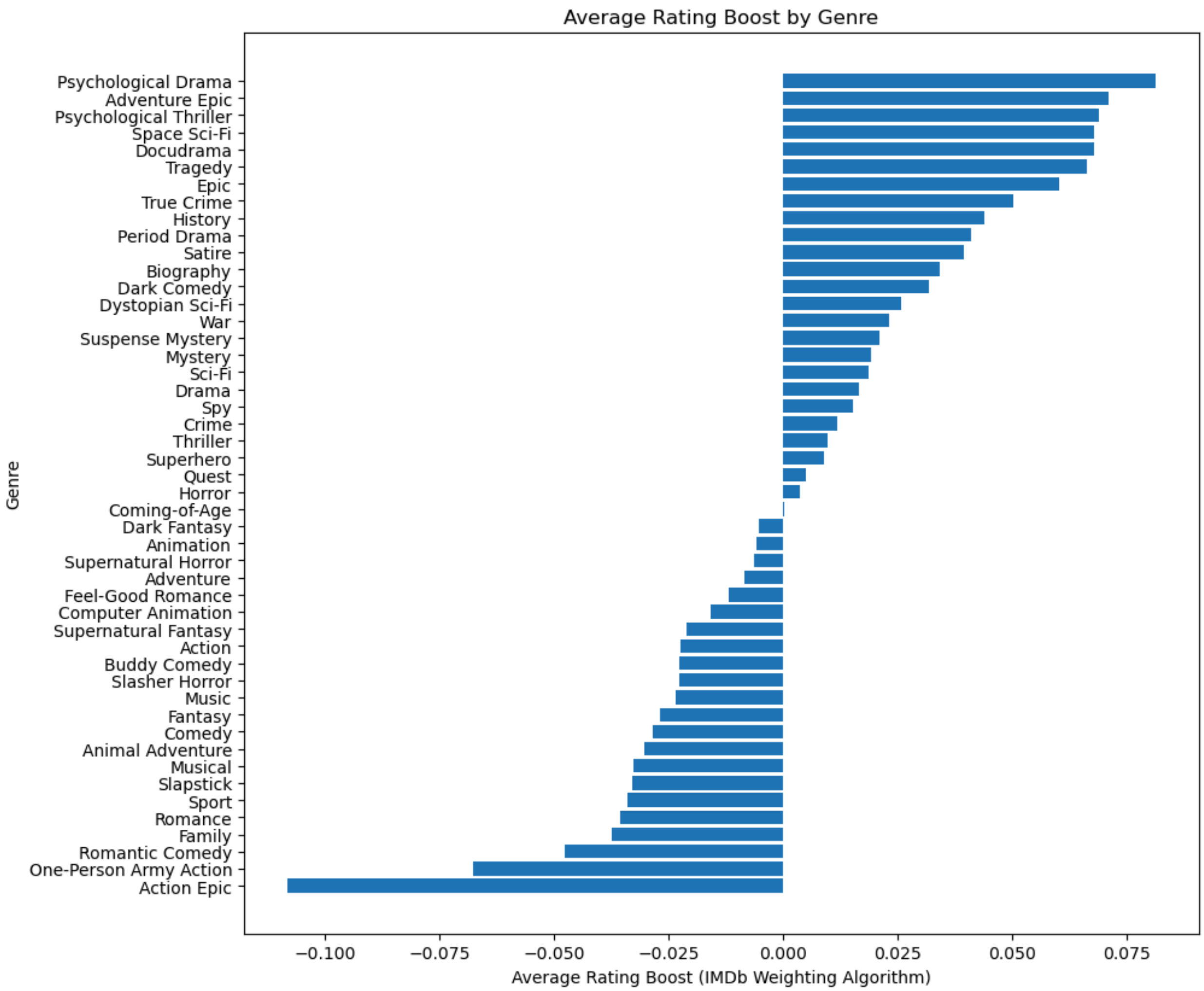
Exploring IMDB's Rating Algorithm



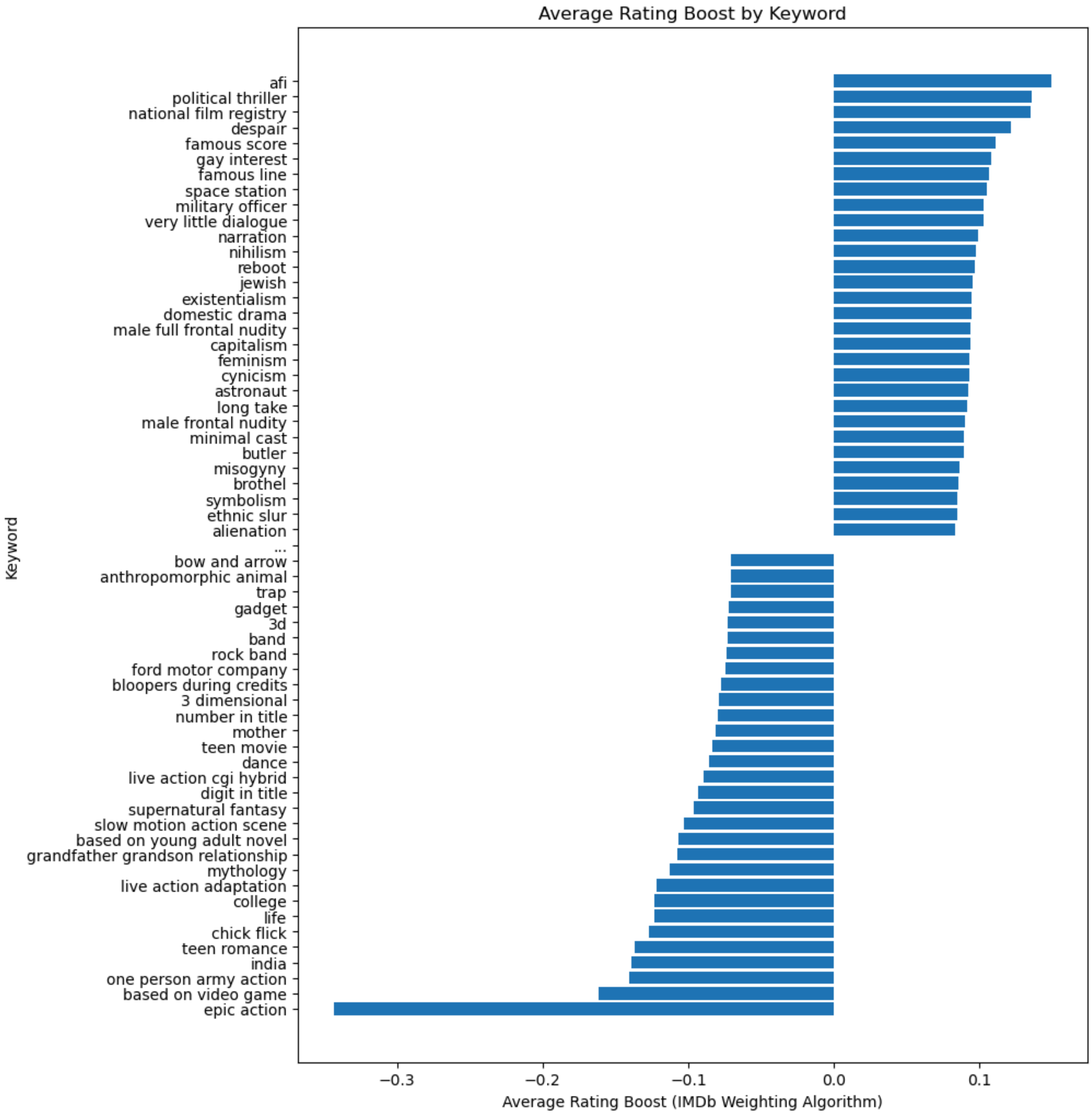
The distribution of rating difference shows that IMDb's weighting algorithm tends to yield lower scores than the unweighted ratings, with some exceptions. Most films' ratings are adjusted by tenths of a point or less, although some, visible on the ends of the distribution, have their ratings drastically altered by the algorithm.



In general, it seems as though films with higher polarization scores are more likely to have their ratings heavily adjusted by IMDb's weighting algorithm (positive, albeit weak, correlation between absolute rating difference and polarization score). This supports the idea that the weight algorithm, either by design or not, mitigates the effects of review-bombing, which results in abnormally polarized rating distributions.

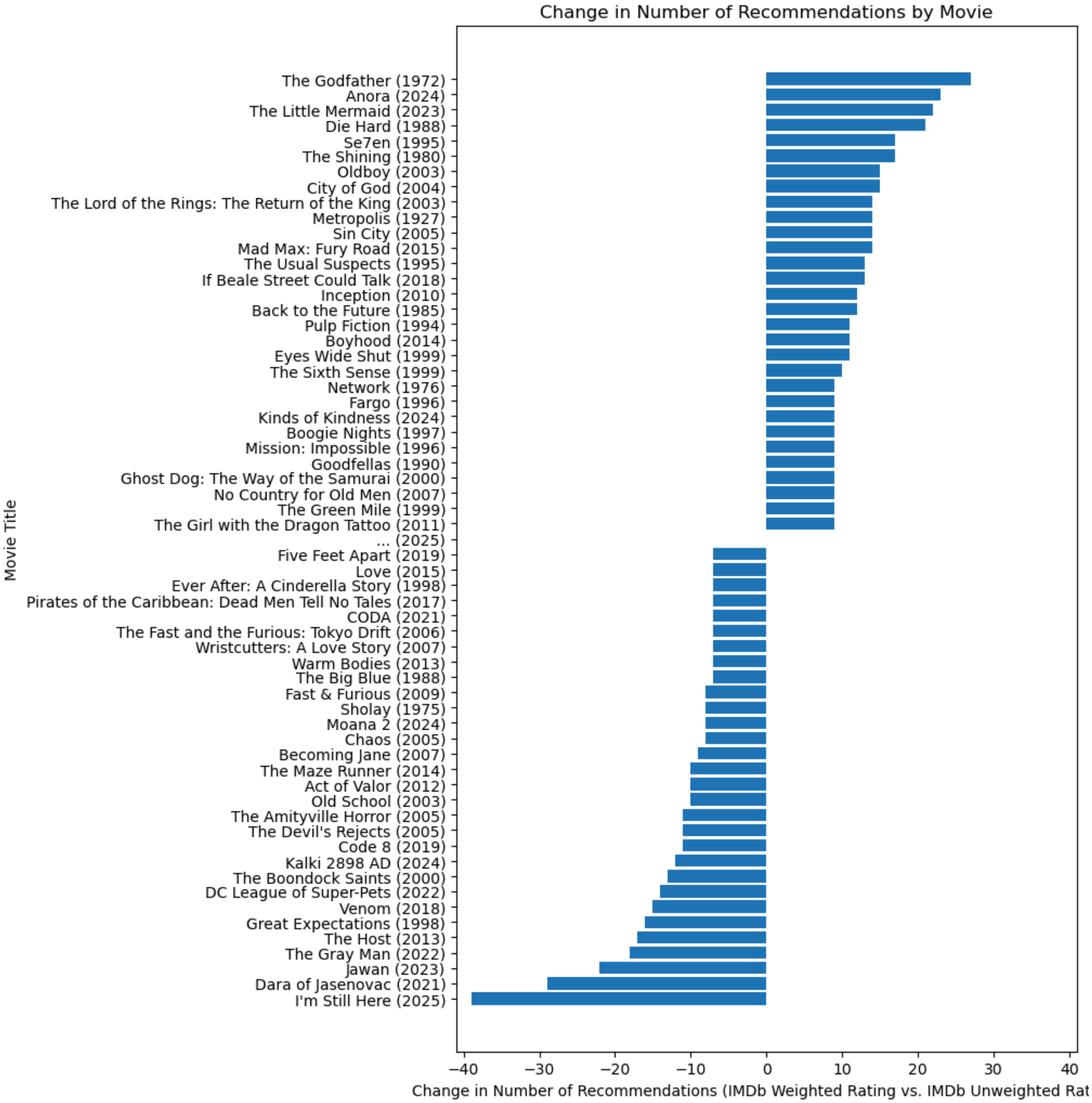


Genres associated with political subject material (e.g. "War", "Period Drama", "Docudrama", "History", "Biography", "Satire") as well as genres more likely to have violent or non-culturally normative themes (e.g. "Psychological Drama", "Psychological Thriller", "True Crime", "Dark Comedy", "Dystopian Sci-fi") are among those boosted most by IMDb's weighting algorithm. Genres associated with more mainstream, family-friendly, or culturally normative themes (e.g. "Family", "Romance", "Musical", "Romantic Drama", "Slapstick", "Sport", "Action Epic", "Animal Adventure", "Comedy", "Fantasy") are among those de-boosted most by the algorithm. This suggests that the algorithm is more likely to boost ratings for films with heavier or more political subject matter, while de-boosting ratings for films with more mainstream or family-friendly themes. This is consistent with the idea that IMDb's weighting algorithm is designed to reduce the influence of unrepresentative user ratings on the overall rating, which may be more likely to be positive for films with more mainstream or family-friendly themes and negative for films with actors from underrepresented groups in starring roles, or films whose themes provoke a large-scale social media backlash.



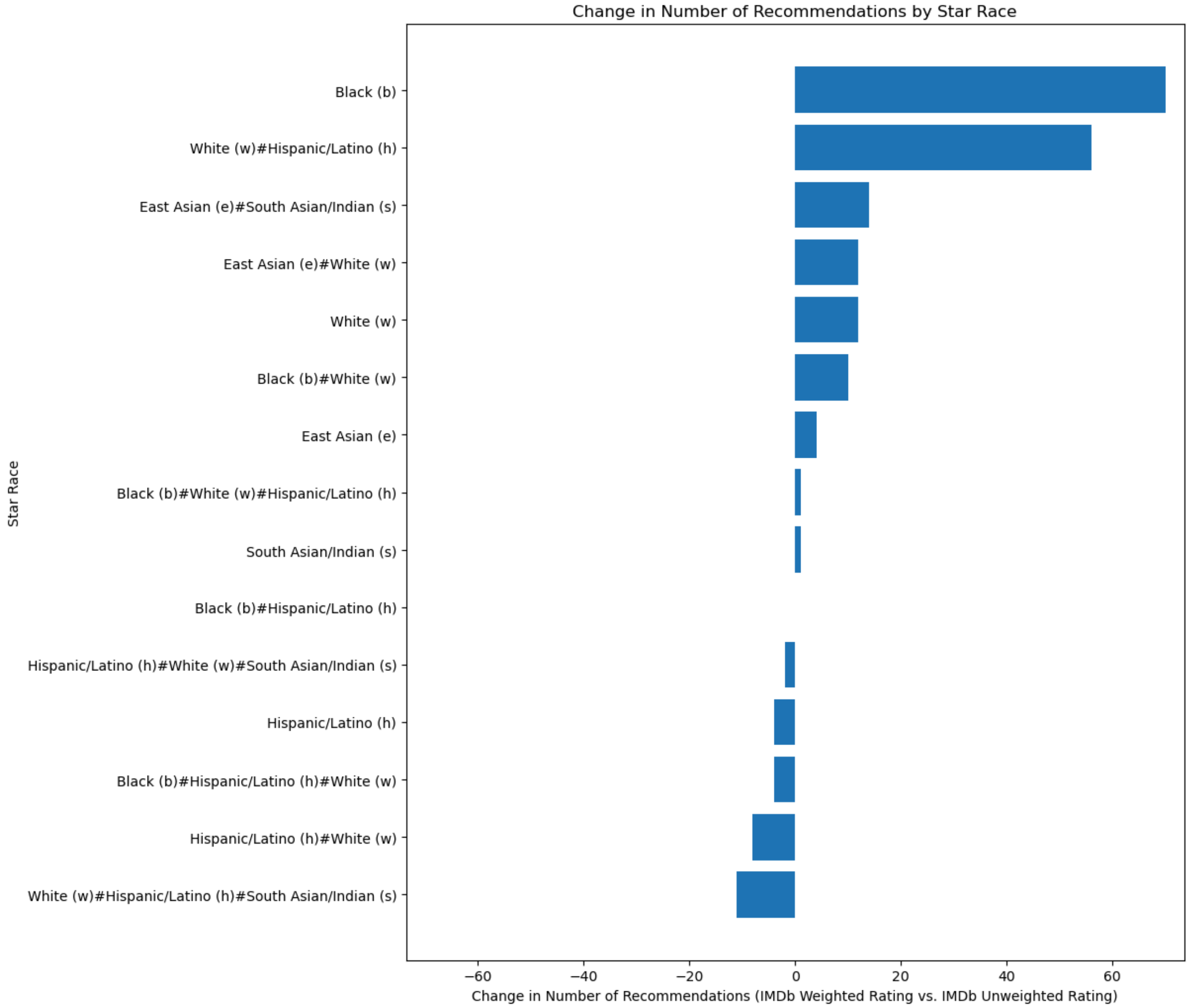
Keywords associated with more high-brow movies (e.g. "afi", "national film registry", "very little dialogue", "symbolism", "minimal cast") as well as keywords associated with political subject material, or keywords relating to marginalized groups or the topic of bigotry itself (e.g. "political thriller", "gay interest", "jewish", "capitalism", "feminism", "male frontal nudity", "misogyny", "brothel", "ethnic slur", "alienation") are among those boosted most by IMDb's weighting algorithm. Keywords associated with more mainstream movies ("epic action", "based on video game", "teen romance", "chick flick", "college", "live action adaptation", "based on young adult novel", "teen movie") are among those de-boosted most by IMDb's weighting algorithm. This matches the observations on how weighting affects different genres.

Effect of IMDb’s Weighting Algorithm on a Hypothetical Movie Recommendation System



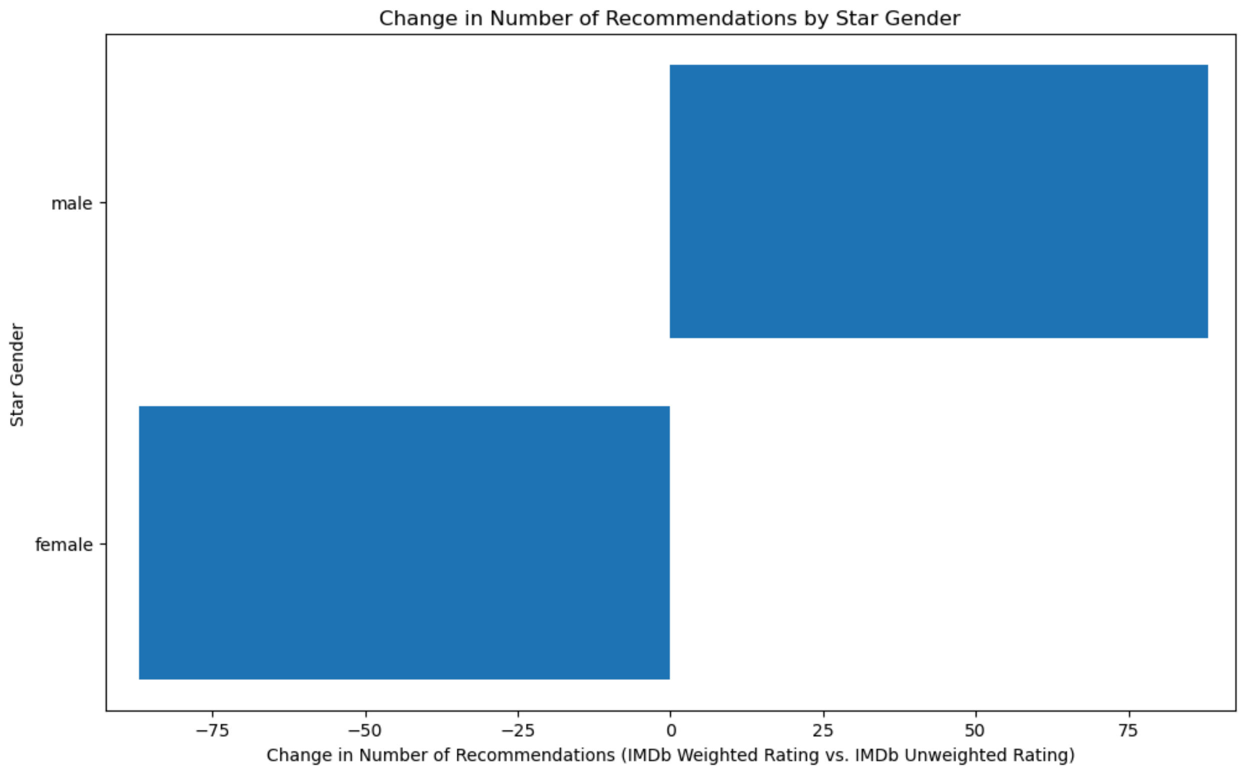
When IMDb’s weighting algorithm is applied, the movies that received an increase in the number of recommendations in our hypothetical movie recommendation system are generally highly rated movies that are not affected considerably by the weighting algorithm. A notable exception is *The Little Mermaid* (2023), which is the movie that received the third highest increase in number of recommendations (22) and also received a large boost from the weighting algorithm (2.07). On the other end, the two movies that received the largest decrease (-29 and -39, respectively) in number of recommendations, *Dara of Jasenovac* (2021) and *I’m Still Here* (2025) are both movies that received a significant de-boost from the weighting algorithm (-1.29

and -0.86, respectively). Both *Dara of Jasenovac* and *I'm Still Here* are political in topic and seem to be recipients of positive review-bombing. These results suggest that one of the effects that IMDb's weighting algorithm has on movie recommendation algorithms is mitigating the effect of review-bombing.



In our hypothetical movie recommendation system, among the top 15 most common race/ethnicity labels in our dataset, Black and Hispanic white stars received the largest increase in number of recommendations (70 and 56, respectively).[4]

[4] One recommendation for a movie was counted as one recommendation for each of its stars.



In our hypothetical movie recommendation system, male stars received an increase in the number of recommendations (88) while female stars received an almost complementary decrease (-87).

DISCUSSION

Based on the results generated by our recommender, we found that movies with keywords such as political thriller, capitalism, feminism, misogyny, etc, were among those boosted most by the IMDB weighting algorithm. We also found that films with more “neutral” keywords (epic action, mythology, college, etc) received a negative rating boost. Out of the ten movies with the highest percentage of polarized votes, we found that the genre that was represented the most was “Political Drama/Thriller.” This is consistent with the idea that IMDb's weighting algorithm is designed to reduce the influence of unrepresentative user ratings on the overall rating, which may be more likely to be positive for films with more mainstream or family-friendly themes and more negative for films with political subject matter or films that received an (often racist or misogynistic) online backlash.

Expanding on the findings of Hickey and Mehta, we gained a clearer understanding of which movie genres and keywords were more commonly associated with instances of review bombing. Moreover, our analysis using the basic recommender allows us to more accurately quantify the possible harms caused by review bombing. However, we also noted that there was not a clear correlation between review bombing and the change in number of recommendations by movie. Thus, at least using our recommender system, review bombing did not have a tangible effect on the visibility of movies in recommender systems.

We also found that cast demographics affected how a movie would receive a recommendation boost or de-boost. Notably, we found that among all the race keywords for a given, movies featuring Black and White-Hispanic/Latino lead actors received some of the highest recommendation boosts. This was in line with our hypothesis, since the recommendation boosts suggests that these movies were negatively review-bombed by reactionary online users. Moreover, we found that gender was a strong factor in review bombing. However, we found that there was not a strong correlation between gender identity and review bombing and re-weighting. Given the online uproar over films like *Ghostbusters* (2016), which featured a predominantly female-identifying cast, we were surprised to see these results.

However, there are a few issues with our dataset. IMDB, while widely used, is not a universal database and does not include all films. In particular, IMDB has fewer titles that are lesser-known or independent. Therefore, our dataset is limited in its scope, making it difficult to make broad generalizations on the effect of review bombing holistically. In addition, our dataset was locked at the time of scraping, so we would have to re-scrape our data in order to analyze the most current movies. In addition, while it helped narrow our dataset, excluding movies with less than 50,000 user votes did somewhat limit the movies we were able to work with. For example, while the documentary *Israelism* (2023) would have been an interesting datapoint to analyze, we ultimately had to exclude it from our dataset based on the vote count criteria.

While our current work is a good starting point, we would like to expand our analysis to include some of the possible far-reaching harms that review bombing can cause. One interesting part of the IMDB dataset is that it includes the top five countries that had the most user engagement with reviews. Thus, we would have also liked to see how region affects the way certain movies are review-bombed. For example, for a movie like *The Promise*, a vast majority of users came from Turkey or Armenia. Thus, we wanted to see if movies that covered certain subjects or had cast members of certain races were more likely to draw reviews from a specific region.

REFLECTIONS

If we had a chance to start the project over, we would have liked to use at least two other recommender systems in addition to the one we used. Given that most streaming services utilize different recommender techniques, we would have liked to incorporate different algorithms into our data analysis to better understand how different recommendation systems are affected differently by review bombing [Hinkle, 2021]. Unfortunately, because streaming sites' recommendation algorithms are black boxes, it is difficult to know exactly what factors are being considered, but a more detailed look at differences in recommendation systems would have been a really interesting direction for the project to go in. Moreover, when looking at the dataset initially, one of the trends that stood out to us most strongly was rating disparities between countries, and we were interested in looking at how the region affected review bombing and seeing which films were the most polarized across international borders. Ultimately, this did not end up becoming a focus, but it could have been the basis of a fascinating analysis, and opened up possibilities with regard to map-based visualizations. Finally, while IMDB is a pretty expansive dataset, we also would have liked to use alternative movie rating sites like MovieLens, Letterboxd, and Rotten Tomatoes, all of which utilize their own form of review weighting. This would have allowed us to compare how effective different weighting systems are at addressing review bombing, opening up another potential avenue for analysis in our work on weighting systems and their effect on recommendations.